Technical critique of a manuscript entitled: ***"Decrease in anogenital distance among male infants with prenatal phthalate exposure"*** published in Environmental Health Perspectives – Online 27 May, 2005 by Swan SH, Main KM, Liu F, Stewart SL, Kruse RL, Calafat AM, Mao CS, Redmon JB, Ternand CL, Sullivan S, Teague JL, Study for Future Families Research Team.

M. Gerald Ott, PhD
BASF Corporation
Corporate Medical Department

Dirk Pallapies, MD, MSc
BASF Aktiengesellschaft
Occupational Medical and Health
Protection Department

8 June 2005

OVERVIEW OF STUDY AND DETAILED COMMENTS

Study

The study by Swan et al (2005) examines statistical associations between physical genital measurements in 85 boys, up to 28 months of age, and a corresponding set of measurements of phthalate monoester metabolites in single spot urine samples collected from their mothers during the pregnancy. Study subjects were recruited from participants in a much larger fertility study of women and their partners at 3 prenatal clinics in the U.S. called the Study for Future Families (SFFI). The purpose of this sub-study was to assess potential adverse reproductive effects (anti-androgenic effects) in the boys relative to phthalate metabolite levels in their mothers. The study is predicated on the observation of anti-androgenic effects in high dose rodent studies with some, but not all, phthalate compounds. These compounds are ubiquitous in the environment and their monoester metabolites have been broadly detected at µg/L concentrations in urine samples collected within the general population.

Comments

This study represents a first attempt at linking physical genital measurements in young boys to indices of environmental chemical exposure in their mothers. Only one other small study has reported findings related to the systematic measurement of such characteristics in young boys (Salazar-Martinez et al., 2004). Thus, there are essentially no historical data or established procedures for performing this study nor a basis for assessing adverse health consequences linked to measurements of anogenital distance in humans. The study results are summarized using categorical and regression analyses that are largely repetitive differing mainly in what transformations of the exposure or health outcome measures were utilized. Very little consideration has been given to assessment of potential confounding bias (that is, bias due to factors that are both correlates of exposure and the health outcomes being reported). There has also been little attention given to assessing possible selection and measurement biases. Overall, the interpretations and conclusions of the study are not well justified. As outlined below, a number of serious methodological and interpretive issues are raised regarding this study.

Methodological Issues

1. **Due to various exclusions of subjects and nonparticipation, the study group may not be representative of the general population. Certain exclusions, in particular, could have introduced bias into the study.**

   The study manuscript refers to an earlier publication (Swan et al, 2003) for a description of the SFFI methods, but does not indicate the participation rate during initial recruitment to the phase I study. It is indicated that 85% of the SFFI participants agreed to be contacted about participation in further studies and these

were the families invited to participate in the follow-up study described as SFFII. However, additional eligibility requirements were imposed that were met by only 73% of the invited participants. One of these requirements was that the mother could attend at least one study visit. The manuscript does not provide information on how many mothers could not meet this requirement. It was then stated that 72.5% of these eligible mothers did participate in the phase II study. Thus, participation was a maximum of 62% (.85 x 72.5%) and may have been much lower. An additional confusing statement is: "Of the 172 boys born to these mothers, five boys in twin births were excluded, leaving 156 boys with data from the first examination." (See page 10 of manuscript). This statement suggests that additional exclusions may have arisen because of incomplete data on the first examination.

Further exclusions (23 of 156 boys or 15%) were reported based on incomplete genital measurement data. Two boys did not have a genital examination (mother refused the examination) and, for 21 boys, the study examiner felt the measurement of anogenital distance was not reliable. The latter exclusion category may be important if the examiners felt that longer or shorter relative distances were more difficult to measure. Information was provided indicating that these boys were on average 3.7 months older than the remaining boys and were usually more active than the other boys. There are no indications as to whether or not the urinary phthalate metabolite levels were different for the mothers of these boys compared to participating boys. This issue could also be relevant to the construction of an anogenital index (AGI) as a function of weight and linear and quadratic age terms as discussed below.

2. The anogenital index (AGI) used as the primary outcome measure in this study is novel and the age-adjusted AGI was constructed artificially and not based on biological rationale.

The AGI was calculated as the distance between the center of the anus and the anterior base of the penis divided by the weight of the boy and was expressed in mm/kg units. This is a novel measure. Previously, a study by Salazar-Martinez (2004) had measured the distance from the center of the anus to the junction of the smooth perineal skin with the rugated skin of the scrotum in male infants (ASD). This parameter was measured in newborn infants rather than infants varying in age from newborn to more than 24 months. In boys, this parameter was found to correlate somewhat better with birth length than birth weight and appears from the diagrams provided to be much easier to measure.

One would expect infant weight, height, and age to correlate with one another. Therefore constructing an age-adjusted AGI, where the distance measure has already been corrected for total infant weight, appears very artificial. Additionally, both age and age squared terms were used in the regression equation to maximize the total variability accounted for. This measure was apparently

favored over the previously published ASD measure, simply because the amount of variability explained by age and age squared terms was somewhat higher for the AGI measure (see page 12 of Swan electronic text).

The AGI index, that adjusts the distance measure using correlated parameters (weight, age, and age squared), could introduce spurious correlations given the small number of participants because it minimizes the remaining between infant variability that is available to be explained by exposure or other factors by using terms such as age squared. This outcome variable has not been researched before and there are no standardized data for the general infant population or extensive assessment of measurement error. It is not at all clear why height of infant was not considered as a standardizing variable. This is especially puzzling given that age was not calculated in the same way for all infants. For those infants with a gestation period of <38 weeks and who were less than 1 year of age at the time of measurement, their age was recalculated from date of conception rather than date of birth. Thus, age is calculated differently for some boys (by up to 8 months) only based on a difference of a few weeks in gestation period. This calculation of age, using two different starting points, could impact all analyses relating age-adjusted AGI to other genital measurements that vary with age.

Standardization of the examination and measurement procedure across the three participating clinics in the Swan et al. study is critical. In the Salazar-Martinez study, where all measurements were performed in a single clinic, information was provided describing the position of the infant during measurement as well as the instrument used in taking the measurements. In the Swan study, it was indicated that "Every attempt was made to standardize the examination …", however, no detail was provided on the position of the baby, the instruments used, or how consistency in measurement was assured across participating clinics. The latter point is particularly germane in that confounding could inadvertently be introduced if phthalate levels differed across the clinics due to geographic factors and there were differential measurement results across clinics as well. Similarly, both AGI measurements and phthalate levels may have varied over calendar time during the three-year data collection period for the study.

3. Reliance on analysis of a single spot urine sample, taken at some time during the pregnancy, as the sole indicator of exposure to phthalate esters throughout the relevant period of pre- and post-natal exposure to the infant is problematical because of the relatively short half-lives of these esters in the body.

Phthalates are rather ubiquitous compounds that may be present in air, food, water, personal care products, and medical devises as well as in some medications. Diethyl phthalate has even been identified in secretions produced by *Helicobacter pylori*, a bacterium commonly infecting human gastric tissue (Keire et al., 2001). These compounds have biological half-lives measured in hours (Schmid and Schlatter, 1985; Hauser et al., 2004), thus, a single urine sample

most often reflects exposure experienced during the preceding day. Monoester phthalate metabolites are measured in urine as indicators of exposure to the parent diester compounds present in the environment. For example, the most abundant phthalate detected in urine is monoethyl phthalate (MEP), a metabolite of diethyl phthalate. Hauser explicitly recommended collecting at least two urine samples 1-3 months apart when assessing male reproductive endpoints, although repeat sampling on 10 subjects indicated that nondifferential random exposure misclassification is likely to be moderate or small for most phthalate metabolites of interest (Hauser et al., 2004). A single urine sample was reported to be most predictable for assessing monoethyl phthalate (MEP), which is not regarded to be a reproductive toxin (ATSDR, 1995).

In the Swan study, mention is made of having analyzed 214 urine samples for phthalate levels including post-natal maternal and baby samples. No explanation is given for why only 85 samples were utilized in assessing phthalate exposures to the infants during the period of development up until the time of their examination.

4. Defining a phthalate exposure score that sums score values for phthalate metabolites with and without evidence of reproductive effects and weighting each metabolite equally regardless of differences in toxic potency or absolute concentration is highly artificial and not biologically based.

A score variable was constructed using 4 of 9 monoester phthalate metabolites that were measured in prenatal urine. A first step was to conduct regression analyses relating AGI to the logarithm of the prenatal monoester metabolite concentration for each metabolite. These results are summarized in Table 4 for 8 of the 9 metabolites discussed in the Swan manuscript. The exclusion of findings for one metabolite (mono-3-carboxypropyl phthalate) in Table 4 may have been inadvertent as it was included in the table footnote. P-values regarding the association between AGI and individual metabolites were then used in selecting the 4 metabolites to be included in the score variable; those with p-values ranging from 0.012 to 0.05 were included. The metabolites with p-values ranging from 0.084 to 0.772 were excluded. Given that each of the 4 selected metabolites was individually associated with lower AGI, it is not surprising that a score variable constructed from an equal weighting of each of the component metabolites would yield a similar association as the individual metabolites. However, the parent compound for the metabolite showing the strongest association (lowest p-value) is not toxicologically linked to male reproductive effects despite extensive testing, while the parent compound of the metabolite with the weakest association (mono-2-ethylhexyl phthalate) is toxicologically linked to male reproductive effects and is regarded as being the most potent in producing reproductive and developmental effects by Koo et al. (2002). Yet this latter metabolite was not included in the score measure.

The scoring system itself is highly artificial, being based on summing the assigned score value across the 4 selected metabolites, where individual metabolite scores were defined based on quartile distributions of results across the 85 participants. The scores assigned were 0, 1, 2, or 3 depending on quartile distribution. This assignment guarantees that each metabolite carries an equal weighting regardless of toxic potency or absolute metabolite level. For example the 75% percentile value for mono-ethyl phthalate is 436.9 ng/mL, whereas the 75% percentile value for the mono-isobutyl phthalate metabolite is 5.1 ng/mL (see Table 3 of manuscript). Comparisons are then made between subjects having cumulative scores of 0-1 (n = 11) versus those having cumulative scores of 11-12 (n = 10). Such comparisons are simply not interpretable.

5. Transformation of continuous variables into categorical variables may create artifacts through arbitrary grouping of exposure or outcome variables.

Because the phthalate ester concentrations observed in general populations tend to be log-normally distributed, sub-categorization according to tertile or quartile may ignore information important in assessing dose-response. Categorization into tertiles makes little biological sense when concentrations were stated to vary by up to 4 orders of magnitude. Such analyses are problematical where higher levels of statistical significance are achieved simply by grouping the data differently.

Results Issues

6. The basis for analyses purportedly demonstrating associations between (1) a short AGI and incompletely descended testes and (2) a short AGI and scrotum categorized as small and/or "not distinct from the surrounding tissue" appear to be poorly described and characterized.

In the Swan manuscript it is stated that: "Among the 134 boys for whom we have genital measurements, no frank genital malformations or disease were detected and no parameters appeared grossly abnormal." Therefore, presumably none of the boys had cryptorchidism, a developmental defect characterized by failure of the testes to descend into the scrotum. In a study of a large number of consecutively recruited births in Denmark and Finland, where testicular position was assessed by a standardized technique and retractile testes were considered normal, the proportion of boys with cryptorchidism depended on age at examination (Boisen et al., 2004). Among Danish boys, the prevalence of cryptorchidism at birth was 9.0% and declined to 1.9% by 3 months of age. Thus, assessment of the status of the testes and presumably the distinctness of scrotal tissue is age-dependent and may depend on other factors such as being small for gestational age and mode of delivery. There is no indication whether or not these factors were taken into consideration when evaluating these categorical

parameters in relation to "short AGI", which was adjusted in some fashion for age.

7. The statistical test regarding the relationship between incompletely descended testes and having short AGI compared to all other boys appears to be in error.

An important argument is made in this paper on page 12 that the proportion of boys with one or both testes incompletely descended is significantly correlated with having short vs. intermediate or long AGI. These findings were used to support the argument in the abstract that the overall findings are consistent with "the phthalate-related syndrome of incomplete virilization". However, it appears that the p-value of < 0.001 was incorrectly computed.

Among the total group of 134 boys with genital measurements, the proportion with incompletely descended testes was 13.4% (see page 10). This indicates that 18 boys in all had incompletely descended testes. There were 85 boys with both genital and phthalate measurements. Among those boys, whose AGI was classified as short (n=24), 20.8% or 5 had incompletely descended testes. Among those boys, with an AGI classified as intermediate (n = 46), 8.9% or 4 had incompletely descended testes and among boys, with an AGI classified as long (n=15), 6.7% or 1 had incompletely descended testes. This yields the following two by two contingency table:

	Incompletely descended	Normal	Total
Short AGI	5	19	24
Other AGI	5	56	61
Total	10	75	85

These proportions (20.8 vs. 8.1%) are not statistically different using a Fisher Exact Test ($p = 0.11$) and the p-value is certainly not < 0.001. Furthermore, the proportion of the remaining boys with incompletely descended testes and without phthalate measurements (8 of 49 boys or 16.3%) is close to that seen in the short AGI group.

Interpretive Issues

8. There is very little discussion of potential confounding factors in the study; these would consist of factors that are correlates of phthalate levels in the studied population and are also linked to health outcomes being assessed.

Potential confounding could arise is several ways. In a general population sample of 289 participants, urinary phthalate levels have been shown to vary by education level, family income level, and urban versus rural residence (Koo et al, 2002). Cryptorchidism has also been shown to vary inversely by education level (Pierik et al., 2004). Thus, education level could easily act as a confounder in assessing the relationship of phthalate levels to incompletely descended testes. The lack of any discussion regarding the distribution of phthalate levels by clinic center is a particular concern, because the patient populations in the three clinics may differ from one another and because the measurement procedures appear not to have been validated across participating clinics. At the very least, clinic should have been treated as a stratifying variable in assessing key associations between exposure and outcome.

9. Some statements concerning the implications of the findings in the discussion section of the report appear to reach well beyond the limitations of the data analyzed.

Various statements in the Swan manuscript appear to ignore or gloss over biological plausibility issues such as whether or not the metabolite concentrations are in a relevant range to potentially produce toxic effects. This is very important to consider given that the study population was found to have phthalate metabolite levels in the expected ranges for the general population. This means that the findings could have immediate consequences if, in fact, these levels are even near what regulatory agencies would consider of toxicological concern.

Stating categorically that "AGI, the most sensitive marker of anti-androgen action in toxicologic studies, is shortened and testicular descent impaired in boys whose mothers had elevated prenatal phthalate exposure" on the basis of the information provided in the manuscript is highly presumptive. Furthermore, the statement that "These changes in male infants, associated with prenatal exposure to some of the same phthalate metabolites that cause similar alterations in male rodents, suggest that commonly used phthalates may undervirilize humans as well as rodents" reaches beyond the limitations of the data analyzed and ignores results from assessments of human reproductive risks related to phthalates as conducted by the National Toxicology Program and reviewed recently by McKee et al.(2004).

References

Agency for Toxic Substances and Disease Registry (ATSDR). 1995. Toxicological profile for diethyl phthalate. Atlanta, GA: U.S. Department of Health and Human Services, Public Health Service.

Boisen KA, Kaleva M, Main KM, et al. 2004. Difference in prevalence of congenital cryptorchidism in infants between two Nordic countries. *Lancet* Apr 17 **363**(9417):1264-1269.

Hauser R, Meeker JD, Park S, et al. 2004. Temporal variability of urinary phthalate metabolite levels in men of reproductive age. *Environ Health Perspect* **112**: 1734-1740.

Keire DA, Anton P, Faull KF, et al. (2001). Diethyl phthalate, a chemotactic factor secreted by Helicobacter pylori. *J Biol Sci* **276**:48847-48853.

Koo JW, Parham F, Kohn MC, et al. 2002. The association between biomarker-based exposure estimates for phthalates and demographic factors in a human reference population. *Environ Health Perspect* **110**:405-410.

McKee RH, Butala JH, David RM, Gans G. 2004. NTP center for the evaluation of risks to human reproduction reports on phthalates: addressing the data gaps. *Reprod Toxicol* **18**:1-22.

Pierik FH, Burdorf A, Deddens JA, et al. 2004. Maternal and paternal risk factors for cryptorchidism and hypospadias: A case-control study in newborn boys. *Environ Health Perspect* **112**:1570-1576.

Salazar-Martinez E, Romano-Riquer P, Yanez-Marquez E, et al. 2004. Anogenital distance in human male and female newborns: a descriptive, cross-sectional study. *Environ Health* **3**:8.

Schmid P, Schlatter C. 1985. Excretion and metabolism of di(2-ethylhexyl)phthalate in man. *Xenobiotica* **15**:251-256.

Swan SH, Brazil C, Drobnis EZ, et al. 2003. Geographic differences in semen quality of fertile U.S. males. *Environ Health Perspect* **111**:414-420.

Swan SH, Main KM, Liu F, et al. 2005. Decrease in anogenital distance among male infants with prenatal phthalate exposure. *Environ Health Perspect* (available at http://dx.doi.org/) Online 27 May 2005.